(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
30 January 2003 (30.01.2003)

PCT

(10) International Publication Number
WO 03/008440 A2

(51) International Patent Classification[7]: C07K 14/00

(21) International Application Number: PCT/EP02/07929

(22) International Filing Date: 16 July 2002 (16.07.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/305,806	16 July 2001 (16.07.2001)	US
60/358,416	20 February 2002 (20.02.2002)	US

(71) Applicant *(for all designated States except US)*: **SYNGENTA PARTICIPATIONS AG** [CH/CH]; Schwarzwaldallee 215, CH-4058 Basel (CH).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **LEVIN, Joshua, Zvi** [US/US]; Syngenta Biotechnology, Inc., 3054 Cornwallis Road, Research Triangle Park, NC 27709 (US). **PATTON, David, Andrew** [US/CH]; Syngenta Crop Protection AG, Werk Stein, Schaffhauserstasse, CH-4332 Stein (CH). **MCELVER, John, Alan** [US/US]; Syngenta Biotechnology, Inc., 3054 Cornwallis Road, Research Triangle Park, NC 27709 (US). **BUDZISZEWSKI, Gregory, Joseph** [US/US]; Syngenta Biotechnology, Inc., 3054 Cornwallis Road, Research Triangle Park, NC 27709 (US). **ZHOU, Qing** [CN/US]; Syngenta Biotechnology, Inc., 3054 Cornwallis Road, Research Triangle Park, NC 27709 (US). **AUX, George, W.** [US/US]; Syngenta Biotechnology, Inc., 3054 Cornwallis Road, Research Triangle Park, NC 27709 (US). **TOSSBERG, John** [US/US]; Syngenta Biotechnology, Inc., 3054 Cornwallis Road, Research Triangle Park, NC 27709 (US). **WEGRICH GLOVER, Lyn** [US/US]; 5446 Escover Lane, San Jose, CA 95118 (US). **ASHBY, Carl, Sandidge** [US/US]; Capital One FSD, 12061-0148, 11013 W. Broad Street, Glen Allen, VA 23060 (US). **THOMAS, Carl, Randall** [US/US]; Syngenta Biotechnology, Inc., 3054 Cornwallis Road, Research Triangle Park, NC 27709 (US). **MADHAVEN, Ernie** [US/US]; Falmouth, VA (US). **LEWIS, Sharon** [US/US]; Syngenta Biotechnology, Inc., 3054 Cornwallis Road, Research Triangle Park, NC 27709 (US). **DUNN, Jill** [US/US]; Syngenta Biotechnology, Inc., 3054 Cornwallis Road, Research Triangle Park, NC 27709 (US). **CATES, Eddie** [US/US]; Syngenta Biotechnology, Inc., 3054 Cornwallis Road, Research Triangle Park, NC 27709 (US). **LAW, Marcus, Dixon** [US/US]; Syngenta Biotechnology, Inc., 3054 Cornwallis Road, Research Triangle Park, NC 27709 (US).

(74) Agent: **BASTIAN, Werner**; Syngenta Participations AG, Intellectual Property, P.O. Box, CH-4002 Basel (CH).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 03/008440 A2

(54) Title: NUCLEIC ACID MOLECULES ENCODING PROTEINS ESSENTIAL FOR PLANT GROWTH AND DEVELOPMENT AND USES THEREOF

(57) Abstract: Nucleotide sequences are isolated from *Arabidopsis thaliana* that code for proteins essential for plant growth and developement. The essentially of the proteins may be exploited by recombinantly expressing the proteins and using them in screening assays to identify compounds that interact with or inhibit the proteins and are therefore potential herbicides.

NUCLEIC ACID MOLECULES ENCODING PROTEINS ESSENTIAL FOR PLANT GROWTH AND DEVELOPMENT AND USES THEREOF

The present invention pertains to nucleic acid molecules isolated from *Arabidopsis thaliana* comprising nucleotide sequences that encode proteins essential for plant growth and development. The invention particularly relates to methods of using these proteins as herbicide targets, based on this essentiality.

The use of herbicides to control undesirable vegetation such as weeds in crop fields has become almost a universal practice. The herbicide market exceeds 15 billion dollars annually. Despite this extensive use, weed control remains a significant and costly problem for farmers.

Effective use of herbicides requires sound management. For instance, the time and method of application and stage of weed plant development are critical to achieving good weed control with herbicides. Because various weed species are resistant to herbicides, the production of effective new herbicides becomes increasingly important. New herbicides can now be discovered using high-throughput screens that implement recombinant DNA technology. Metabolic enzymes found to be essential to plant growth and development can be recombinantly produced through standard molecular biological techniques and utilized as herbicide targets in screens for novel inhibitors of the enzyme activity. More generally, any essential plant protein can be used to screen for inhibitors of its activity. The novel inhibitors discovered through such screens may then be used as herbicides to control undesirable vegetation.

In view of the above, there remain persistent and ongoing problems with unwanted or detrimental vegetation growth (*e.g.* weeds). Furthermore, as the population continues to grow, there will be increasing food shortages. Therefore, there exists a long felt, yet unfulfilled need, to find new, effective, and economic herbicides.

In view of these needs, it is an object of the invention to provide nucleic acid molecules from *Arabidopsis thaliana* comprising nucleotide sequences that encode proteins essential for plant growth and development. It is another object to provide the essential proteins encoded by these essential nucleotide sequences for assay development to identify

inhibitory compounds with herbicidal activity. It is still another object of the present invention to provide an effective and beneficial method for identifying new or improved herbicides using the essential proteins of the invention.

In furtherance of these and other objects, the present invention provides nucleic acid molecules isolated from *Arabidopsis thaliana* comprising nucleotide sequences that encode proteins essential for plant viability. Genetic results show that when any of the nucleotide sequences of the invention are mutated in *Arabidopsis thaliana*, the resulting phenotype is embryo or seedling lethal in the homozygous state. In particular, by using *Ac/Ds* transposon or T-DNA-mediated mutagenesis, the inventors of the present invention are the first to demonstrate that the activity of each protein of the present invention is essential for plant growth in *Arabidopsis thaliana*.

This knowledge is exploited to provide novel herbicide modes of action. The critical role in plant growth of the proteins encoded by each of the nucleotide sequences of the invention implies that chemicals that inhibit the function of any one of these proteins in plants are likely to have detrimental effects on plants and are potentially good herbicide candidates. Thus, the proteins encoded by the essential nucleotide sequences provide the bases for assays designed to easily and rapidly identify novel herbicides.

The present invention therefore provides methods of using a purified protein encoded by any one of the nucleotide sequences described below to identify inhibitors thereof, which can then be used as herbicides to suppress the growth of undesirable vegetation, *e.g.* in fields where crops are grown, particularly agronomically important crops such as maize and other cereal crops such as wheat, oats, rye, sorghum, rice, barley, millet, turf and forage grasses, and the like, as well as cotton, sugar cane, sugar beet, oilseed rape, and soybeans.

Disclosed herein are nucleic acid molecules isolated from *Arabidopsis thaliana*. In one embodiment, the present invention provides an isolated nucleic acid molecule comprising a nucleotide sequence, the complement of which hybridizes under stringent conditions to a sequence selected from the group consisting of the odd numbered SEQ ID NOs:1-95. In another embodiment, the present invention provides an isolated nucleic acid molecule comprising a nucleotide sequence that encodes a protein comprising an amino acid sequence having at least 60%, preferably 70%, more preferably 80%, still more preferably 90%, even more preferably 95%, and most preferably 99-100% sequence identity to an amino acid sequence selected from the group consisting of the even numbered SEQ ID NOs:2-96.

The present invention also provides a chimeric construct comprising a promoter operatively linked to a nucleic acid molecule according to the present invention, wherein the promoter is preferably functional in a eukaryote, wherein the promoter is preferably heterologous to the nucleic acid molecule. The present invention further provides a recombinant vector comprising a chimeric construct according to the present invention, wherein said vector is capable of being stably transformed into a host cell. The present invention still further provides a host cell comprising a nucleic acid molecule according to the present invention, wherein said nucleic acid molecule is preferably expressible in the cell. The host cell is preferably selected from the group consisting of a plant cell, a yeast cell, an insect cell, and a prokaryotic cell. The present invention additionally provides a plant or seed comprising a plant cell according to the present invention.

The present invention also provides proteins essential for plant growth in *Arabidopsis thaliana*. In one embodiment, the present invention provides an isolated protein comprising an amino acid sequence having at least 60%, preferably 70%, more preferably 80%, still more preferably 90%, even more preferably 95%, and most preferably 99-100% sequence identity to an amino acid sequence selected from the group consisting of the even numbered SEQ ID NOs:2-96. In accordance with another embodiment, the present invention also relates to the recombinant production of proteins of the invention and methods of using the proteins of the invention in assays for identifying compounds that interact with the protein.

According to another aspect, the present invention provides a method of identifying a herbicidal compound, comprising: (a) combining a polypeptide comprising an amino acid sequence at least 90% identical to an amino acid sequence selected from the group consisting of the even numbered SEQ ID NOs:2-96 with a compound to be tested for the ability to bind to said polypeptide, under conditions conducive to binding; (b) selecting a compound identified in (a) that binds to said polypeptide; (c) applying a compound selected in (b) to a plant to test for herbicidal activity; and (d) selecting a compound identified in (c) that has herbicidal activity. Preferably, the polypeptide comprises an amino acid sequence at least 95% identical to an amino acid sequence selected from the group consisting of the even numbered SEQ ID NOs:2-96. More preferably, the polypeptide comprises an amino acid sequence at least 99% identical to an amino acid sequence selected from the group consisting of the even numbered SEQ ID NOs:2-96. Most preferably, the polypeptide comprises an amino acid sequence selected from the group consisting of the even numbered SEQ ID

NOs:2-96. The present invention also provides a method for killing or inhibiting the growth or viability of a plant, comprising applying to the plant a herbicidal compound identified according to this method.

According to yet another aspect, the present invention provides a method of identifying a herbicidal compound, comprising: (a) combining a polypeptide comprising an amino acid sequence at least 90% identical to an amino acid sequence selected from the group consisting of the even numbered SEQ ID NOs:2-96 with a compound to be tested for the ability to inhibit the activity of said polypeptide, under conditions conducive to inhibition; (b) selecting a compound identified in (a) that inhibits the activity of said polypeptide; (c) applying a compound selected in (b) to a plant to test for herbicidal activity; and (d) selecting a compound identified in (c) that has herbicidal activity. Preferably, the polypeptide comprises an amino acid sequence at least 95% identical to an amino acid sequence selected from the group consisting of the even numbered SEQ ID NOs:2-96. More preferably, the polypeptide comprises an amino acid sequence at least 99% identical to an amino acid sequence selected from the group consisting of the even numbered SEQ ID NOs:2-96. Most preferably, the polypeptide comprises an amino acid sequence selected from the group consisting of the even numbered SEQ ID NOs:2-96. The present invention also provides a method for killing or inhibiting the growth or viability of a plant, comprising applying to the plant a herbicidal compound identified according to this method.

The present invention still further provides a method for killing or inhibiting the growth or viability of a plant, comprising inhibiting expression in said plant of a protein having at least 60%, preferably 70%, more preferably 80%, still more preferably 90%, even more preferably 95%, and most preferably 99-100% sequence identity to an amino acid sequence selected from the group consisting of the even numbered SEQ ID NOs:2-96.

Other objects and advantages of the present invention will become apparent to those skilled in the art and from a study of the following description of the invention and non-limiting examples. The entire contents of all publications mentioned herein are hereby incorporated by reference.

BRIEF DESCRIPTION OF THE SEQUENCES IN THE SEQUENCE LISTING

Odd numbered SEQ ID NOs:1-95 are nucleotide sequences isolated from *Arabidopsis thaliana* that are more fully described in Table 5 below.

Even numbered SEQ ID NOs:2-96 are protein sequences encoded by the immediately preceding nucleotide sequence, *e.g.*, SEQ ID NO:2 is the protein encoded by the nucleotide sequence of SEQ ID NO:1, SEQ ID NO:4 is the protein encoded by the nucleotide sequence of SEQ ID NO:3, etc.

SEQ ID NOs:101-125 are PCR primers.

DEFINITIONS

For clarity, certain terms used in the specification are defined and presented as follows:

"Associated with / operatively linked" refer to two nucleic acid sequences that are related physically or functionally. For example, a promoter or regulatory DNA sequence is said to be "associated with" a DNA sequence that codes for an RNA or a protein if the two sequences are operatively linked, or situated such that the regulator DNA sequence will affect the expression level of the coding or structural DNA sequence.

A "chimeric construct" is a recombinant nucleic acid sequence in which a promoter or regulatory nucleic acid sequence is operatively linked to, or associated with, a nucleic acid sequence that codes for an mRNA or which is expressed as a protein, such that the regulatory nucleic acid sequence is able to regulate transcription or expression of the associated nucleic acid sequence. The regulatory nucleic acid sequence of the chimeric construct is not normally operatively linked to the associated nucleic acid sequence as found in nature.

Co-factor: natural reactant, such as an organic molecule or a metal ion, required in an enzyme-catalyzed reaction. A co-factor is *e.g.* NAD(P), riboflavin (including FAD and FMN), folate, molybdopterin, thiamin, biotin, lipoic acid, pantothenic acid and coenzyme A, S-adenosylmethionine, pyridoxal phosphate, ubiquinone, menaquinone. Optionally, a co-factor can be regenerated and reused.

A "coding sequence" is a nucleic acid sequence that is transcribed into RNA such as mRNA, rRNA, tRNA, snRNA, sense RNA or antisense RNA. Preferably the RNA is then translated in an organism to produce a protein.

Complementary: "complementary" refers to two nucleotide sequences that comprise antiparallel nucleotide sequences capable of pairing with one another upon formation of hydrogen bonds between the complementary base residues in the antiparallel nucleotide sequences.

Enzyme activity: means herein the ability of an enzyme to catalyze the conversion of a substrate into a product. A substrate for the enzyme comprises the natural substrate of the enzyme but also comprises analogues of the natural substrate, which can also be converted, by the enzyme into a product or into an analogue of a product. The activity of the enzyme is measured for example by determining the amount of product in the reaction after a certain period of time, or by determining the amount of substrate remaining in the reaction mixture after a certain period of time. The activity of the enzyme is also measured by determining the amount of an unused co-factor of the reaction remaining in the reaction mixture after a certain period of time or by determining the amount of used co-factor in the reaction mixture after a certain period of time. The activity of the enzyme is also measured by determining the amount of a donor of free energy or energy-rich molecule (*e.g.* ATP, phosphoenolpyruvate, acetyl phosphate or phosphocreatine) remaining in the reaction mixture after a certain period of time or by determining the amount of a used donor of free energy or energy-rich molecule (*e.g.* ADP, pyruvate, acetate or creatine) in the reaction mixture after a certain period of time.

Essential: an "essential" *Arabidopsis thaliana* nucleotide sequence is a nucleotide sequence encoding a protein such as *e.g.* a biosynthetic enzyme, receptor, signal transduction protein, structural gene product, or transport protein that is essential to the growth or survival of the plant.

Expression Cassette: "Expression cassette" as used herein means a nucleic acid molecule capable of directing expression of a particular nucleotide sequence in an appropriate host cell, comprising a promoter operatively linked to the nucleotide sequence of interest which is operatively linked to termination signals. It also typically comprises sequences required for proper translation of the nucleotide sequence. The coding region usually codes for a protein of interest but may also code for a functional RNA of interest, for example antisense RNA or a nontranslated RNA, in the sense or antisense direction. The expression cassette comprising the nucleotide sequence of interest may be chimeric, meaning that at least one of its components is heterologous with respect to at least one of its other components. The expression cassette may also be one that is naturally occurring but has been obtained in a recombinant form useful for heterologous expression. Typically, however, the expression cassette is heterologous with respect to the host, *i.e.*, the particular DNA sequence of the expression cassette does not occur naturally in the host cell and must have been introduced into the host cell or an ancestor of the host cell by a transformation event. The expression of

the nucleotide sequence in the expression cassette may be under the control of a constitutive promoter or of an inducible promoter that initiates transcription only when the host cell is exposed to some particular external stimulus. In the case of a multicellular organism, such as a plant, the promoter can also be specific to a particular tissue or organ or stage of development.

Gene: the term "gene" is used broadly to refer to any segment of DNA associated with a biological function. Thus, genes include coding sequences and/or the regulatory sequences required for their expression. Genes also include nonexpressed DNA segments that, for example, form recognition sequences for other proteins. Genes can be obtained from a variety of sources, including cloning from a source of interest or synthesizing from known or predicted sequence information, and may include sequences designed to have desired parameters.

Heterologous/exogenous: The terms "heterologous" and "exogenous" when used herein to refer to a nucleic acid sequence (*e.g.* a DNA sequence) or a gene, refer to a sequence that originates from a source foreign to the particular host cell or, if from the same source, is modified from its original form. Thus, a heterologous gene in a host cell includes a gene that is endogenous to the particular host cell but has been modified through, for example, the use of DNA shuffling. The terms also include non-naturally occurring multiple copies of a naturally occurring DNA sequence. Thus, the terms refer to a DNA segment that is foreign or heterologous to the cell, or homologous to the cell but in a position within the host cell nucleic acid in which the element is not ordinarily found. Exogenous DNA segments are expressed to yield exogenous polypeptides.

A "homologous" nucleic acid (*e.g.* DNA) sequence is a nucleic acid (*e.g.* DNA) sequence naturally associated with a host cell into which it is introduced.

Hybridization: The phrase "hybridizing specifically to" refers to the binding, duplexing, or hybridizing of a molecule only to a particular nucleotide sequence under stringent conditions when that sequence is present in a complex mixture (*e.g.*, total cellular) DNA or RNA. "Bind(s) substantially" refers to complementary hybridization between a probe nucleic acid and a target nucleic acid and embraces minor mismatches that can be accommodated by reducing the stringency of the hybridization media to achieve the desired detection of the target nucleic acid sequence.

Inhibitor: a chemical substance that inactivates the enzymatic activity of a protein such as a biosynthetic enzyme, receptor, signal transduction protein, structural gene product, or transport protein. The term "herbicide" (or "herbicidal compound") is used herein to define an inhibitor applied to a plant at any stage of development, whereby the herbicide inhibits the growth of the plant or kills the plant.

Interaction: quality or state of mutual action such that the effectiveness or toxicity of one protein or compound on another protein is inhibitory (antagonists) or enhancing (agonists).

A nucleic acid sequence is "isocoding with" a reference nucleic acid sequence when the nucleic acid sequence encodes a polypeptide having the same amino acid sequence as the polypeptide encoded by the reference nucleic acid sequence.

Isogenic: plants that are genetically identical, except that they may differ by the presence or absence of a heterologous DNA sequence.

Isolated: in the context of the present invention, an isolated DNA molecule or an isolated enzyme is a DNA molecule or enzyme that, by the hand of man, exists apart from its native environment and is therefore not a product of nature. An isolated DNA molecule or enzyme may exist in a purified form or may exist in a non-native environment such as, for example, in a transgenic host cell.

Mature protein: protein from which the transit peptide, signal peptide, and/or propeptide portions have been removed.

Minimal Promoter: the smallest piece of a promoter, such as a TATA element, that can support any transcription. A minimal promoter typically has greatly reduced promoter activity in the absence of upstream activation. In the presence of a suitable transcription factor, the minimal promoter functions to permit transcription.

Modified Enzyme Activity: enzyme activity different from that which naturally occurs in a plant (*i.e.* enzyme activity that occurs naturally in the absence of direct or indirect manipulation of such activity by man), which is tolerant to inhibitors that inhibit the naturally occurring enzyme activity.

Native: refers to a gene that is present in the genome of an untransformed plant cell.

Naturally occurring: the term "naturally occurring" is used to describe an object that can be found in nature as distinct from being artificially produced by man. For example, a protein or nucleotide sequence present in an organism (including a virus), which can be

isolated from a source in nature and which has not been intentionally modified by man in the laboratory, is naturally occurring.

Nucleic acid: the term "nucleic acid" refers to deoxyribonucleotides or ribonucleotides and polymers thereof in either single- or double-stranded form. Unless specifically limited, the term encompasses nucleic acids containing known analogues of natural nucleotides which have similar binding properties as the reference nucleic acid and are metabolized in a manner similar to naturally occurring nucleotides. Unless otherwise indicated, a particular nucleic acid sequence also implicitly encompasses conservatively modified variants thereof *(e.g.* degenerate codon substitutions) and complementary sequences and as well as the sequence explicitly indicated. Specifically, degenerate codon substitutions may be achieved by generating sequences in which the third position of one or more selected (or all) codons is substituted with mixed-base and/or deoxyinosine residues (Batzer *et al., Nucleic Acid Res.* 19: 5081 (1991); Ohtsuka *et al., J. Biol. Chem.* 260: 2605-2608 (1985); Rossolini *et al., Mol. Cell. Probes* 8: 91-98 (1994)). The terms "nucleic acid" or "nucleic acid sequence" may also be used interchangeably with gene, cDNA, and mRNA encoded by a gene.

"ORF" means open reading frame.

Percent identity: the phrases "percent identical" or "percent identical," in the context of two nucleic acid or protein sequences, refers to two or more sequences or subsequences that have for example 60%, preferably 70%, more preferably 80%, still more preferably 90%, even more preferably 95%, and most preferably at least 99% nucleotide or amino acid residue identity, when compared and aligned for maximum correspondence, as measured using one of the following sequence comparison algorithms or by visual inspection. Preferably, the percent identity exists over a region of the sequences that is at least about 50 residues in length, more preferably over a region of at least about 100 residues, and most preferably the percent identity exists over at least about 150 residues. In an especially preferred embodiment, the percent identity exists over the entire length of the coding regions.

For sequence comparison, typically one sequence acts as a reference sequence to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are input into a computer, subsequence coordinates are designated if necessary, and sequence algorithm program parameters are designated. The sequence comparison algorithm then calculates the percent sequence identity for the test sequence(s) relative to the reference sequence, based on the designated program parameters.

Optimal alignment of sequences for comparison can be conducted, *e.g.*, by the local homology algorithm of Smith & Waterman, *Adv. Appl. Math.* 2: 482 (1981), by the homology alignment algorithm of Needleman & Wunsch, *J. Mol. Biol.* 48: 443 (1970), by the search for similarity method of Pearson & Lipman, *Proc. Nat'l. Acad. Sci. USA* 85: 2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by visual inspection (*see generally,* Ausubel *et al., infra).*

One example of an algorithm that is suitable for determining percent sequence identity and sequence similarity is the BLAST algorithm, which is described in Altschul *et al., J. Mol. Biol.* 215: 403-410 (1990). Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul *et al.*, 1990). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always > 0) and N (penalty score for mismatching residues; always < 0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when the cumulative alignment score falls off by the quantity X from its maximum achieved value, the cumulative score goes to zero or below due to the accumulation of one or more negative-scoring residue alignments, or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, an expectation (E) of 10, a cutoff of 100, M=5, N=-4, and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix *(see* Henikoff & Henikoff, *Proc. Natl. Acad. Sci. USA* 89: 10915 (1989)).

In addition to calculating percent sequence identity, the BLAST algorithm also performs a statistical analysis of the similarity between two sequences *(see, e.g.,* Karlin &

Altschul, *Proc. Nat'l. Acad. Sci. USA* 90: 5873-5787 (1993)). One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a test nucleic acid sequence is considered similar to a reference sequence if the smallest sum probability in a comparison of the test nucleic acid sequence to the reference nucleic acid sequence is less than about 0.1, more preferably less than about 0.01, and most preferably less than about 0.001.

Pre-protein: protein that is normally targeted to a cellular organelle, such as a chloroplast, and still comprises its native transit peptide.

Purified: the term "purified," when applied to a nucleic acid or protein, denotes that the nucleic acid or protein is essentially free of other cellular components with which it is associated in the natural state. It is preferably in a homogeneous state although it can be in either a dry or aqueous solution. Purity and homogeneity are typically determined using analytical chemistry techniques such as polyacrylamide gel electrophoresis or high performance liquid chromatography. A protein that is the predominant species present in a preparation is substantially purified. The term "purified" denotes that a nucleic acid or protein gives rise to essentially one band in an electrophoretic gel. Particularly, it means that the nucleic acid or protein is at least about 50% pure, more preferably at least about 85% pure, and most preferably at least about 99% pure.

Two nucleic acids are "recombined" when sequences from each of the two nucleic acids are combined in a progeny nucleic acid. Two sequences are "directly" recombined when both of the nucleic acids are substrates for recombination. Two sequences are "indirectly recombined" when the sequences are recombined using an intermediate such as a cross-over oligonucleotide. For indirect recombination, no more than one of the sequences is an actual substrate for recombination, and in some cases, neither sequence is a substrate for recombination.

"Regulatory elements" refer to sequences involved in controlling the expression of a nucleotide sequence. Regulatory elements comprise a promoter operatively linked to the nucleotide sequence of interest and termination signals. They also typically encompass sequences required for proper translation of the nucleotide sequence.

Significant Increase: an increase in enzymatic activity that is larger than the margin of error inherent in the measurement technique, preferably an increase by about 2-fold or greater

of the activity of the wild-type enzyme in the presence of the inhibitor, more preferably an increase by about 5-fold or greater, and most preferably an increase by about 10-fold or greater.

Significantly less: means that the amount of a product of an enzymatic reaction is reduced by more than the margin of error inherent in the measurement technique, preferably a decrease by about 2-fold or greater of the activity of the wild-type enzyme in the absence of the inhibitor, more preferably an decrease by about 5-fold or greater, and most preferably an decrease by about 10-fold or greater.

Specific Binding/Immunological Cross-Reactivity: An indication that two nucleic acid sequences or proteins are substantially identical is that the protein encoded by the first nucleic acid is immunologically cross reactive with, or specifically binds to, the protein encoded by the second nucleic acid. Thus, a protein is typically substantially identical to a second protein, for example, where the two proteins differ only by conservative substitutions. The phrase "specifically (or selectively) binds to an antibody," or "specifically (or selectively) immunoreactive with," when referring to a protein or peptide, refers to a binding reaction which is determinative of the presence of the protein in the presence of a heterogeneous population of proteins and other biologics. Thus, under designated immunoassay conditions, the specified antibodies bind to a particular protein and do not bind in a significant amount to other proteins present in the sample. Specific binding to an antibody under such conditions may require an antibody that is selected for its specificity for a particular protein. For example, antibodies raised to the protein with the amino acid sequence encoded by any of the nucleic acid sequences of the invention can be selected to obtain antibodies specifically immunoreactive with that protein and not with other proteins except for polymorphic variants. A variety of immunoassay formats may be used to select antibodies specifically immunoreactive with a particular protein. For example, solid-phase ELISA immunoassays, Western blots, or immunohistochemistry are routinely used to select monoclonal antibodies specifically immunoreactive with a protein. See Harlow and Lane (1988) *Antibodies, A Laboratory Manual,* Cold Spring Harbor Publications, New York "Harlow and Lane"), for a description of immunoassay formats and conditions that can be used to determine specific immunoreactivity. Typically a specific or selective reaction will be at least twice background signal or noise and more typically more than 10 to 100 times background.

"Stringent hybridization conditions" and "stringent hybridization wash conditions" in the context of nucleic acid hybridization experiments such as Southern and Northern hybridizations are sequence dependent, and are different under different environmental parameters. Longer sequences hybridize specifically at higher temperatures. An extensive guide to the hybridization of nucleic acids is found in Tijssen (1993) *Laboratory Techniques in Biochemistry and Molecular Biology-Hybridization with Nucleic* Acid Probes part I chapter 2 "Overview of principles of hybridization and the strategy of nucleic acid probe assays" Elsevier, New York. Generally, highly stringent hybridization and wash conditions are selected to be about 5°C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH. Typically, under "stringent conditions" a probe will hybridize to its target subsequence, but to no other sequences.

The T_m is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. Very stringent conditions are selected to be equal to the T_m for a particular probe. An example of stringent hybridization conditions for hybridization of complementary nucleic acids which have more than 100 complementary residues on a filter in a Southern or northern blot is 50% formamide with 1 mg of heparin at 42°C, with the hybridization being carried out overnight. An example of highly stringent wash conditions is 0.1 5M NaCl at 72°C for about 15 minutes. An example of stringent wash conditions is a 0.2x SSC wash at 65°C for 15 minutes (*see,* Sambrook, *infra,* for a description of SSC buffer). Often, a high stringency wash is preceded by a low stringency wash to remove background probe signal. An example medium stringency wash for a duplex of, *e.g.*, more than 100 nucleotides, is 1x SSC at 45°C for 15 minutes. An example low stringency wash for a duplex of, *e.g.*, more than 100 nucleotides, is 4-6x SSC at 40°C for 15 minutes. For short probes (*e.g.,* about 10 to 50 nucleotides), stringent conditions typically involve salt concentrations of less than about 1.0 M Na ion, typically about 0.01 to 1.0 M Na ion concentration (or other salts) at pH 7.0 to 8.3, and the temperature is typically at least about 30°C. Stringent conditions can also be achieved with the addition of destabilizing agents such as formamide. In general, a signal to noise ratio of 2x (or higher) than that observed for an unrelated probe in the particular hybridization assay indicates detection of a specific hybridization. Nucleic acids that do not hybridize to each other under stringent conditions are still substantially identical if the proteins that they encode are substantially

identical. This occurs, *e.g.*, when a copy of a nucleic acid is created using the maximum codon degeneracy permitted by the genetic code.

The following are examples of sets of hybridization/wash conditions that may be used to clone nucleotide sequences that are homologues of reference nucleotide sequences of the present invention: a reference nucleotide sequence preferably hybridizes to the reference nucleotide sequence in 7% sodium dodecyl sulfate (SDS), 0.5 M $NaPO_4$, 1 mM EDTA at 50°C with washing in 2X SSC, 0.1% SDS at 50°C, more desirably in 7% sodium dodecyl sulfate (SDS), 0.5 M $NaPO_4$, 1 mM EDTA at 50°C with washing in 1X SSC, 0.1% SDS at 50°C, more desirably still in 7% sodium dodecyl sulfate (SDS), 0.5 M $NaPO_4$, 1 mM EDTA at 50°C with washing in 0.5X SSC, 0.1% SDS at 50°C, preferably in 7% sodium dodecyl sulfate (SDS), 0.5 M $NaPO_4$, 1 mM EDTA at 50°C with washing in 0.1X SSC, 0.1% SDS at 50°C, more preferably in 7% sodium dodecyl sulfate (SDS), 0.5 M $NaPO_4$, 1 mM EDTA at 50°C with washing in 0.1X SSC, 0.1% SDS at 65°C.

A "subsequence" refers to a sequence of nucleic acids or amino acids that comprise a part of a longer sequence of nucleic acids or amino acids (*e.g.*, protein) respectively.

Substrate: a substrate is the molecule that an enzyme naturally recognizes and converts to a product in the biochemical pathway in which the enzyme naturally carries out its function, or is a modified version of the molecule, which is also recognized by the enzyme and is converted by the enzyme to a product in an enzymatic reaction similar to the naturally-occurring reaction.

Transformation: a process for introducing heterologous DNA into a plant cell, plant tissue, or plant. Transformed plant cells, plant tissue, or plants are understood to encompass not only the end product of a transformation process, but also transgenic progeny thereof.

"Transformed," "transgenic," and "recombinant" refer to a host organism such as a bacterium or a plant into which a heterologous nucleic acid molecule has been introduced. The nucleic acid molecule can be stably integrated into the genome of the host or the nucleic acid molecule can also be present as an extrachromosomal molecule. Such an extrachromosomal molecule can be auto-replicating. Transformed cells, tissues, or plants are understood to encompass not only the end product of a transformation process, but also transgenic progeny thereof. A "non-transformed," "non-transgenic," or "non-recombinant" host refers to a wild-type organism, *e.g.*, a bacterium or plant, which does not contain the heterologous nucleic acid molecule.

Viability: "viability" as used herein refers to a fitness parameter of a plant. Plants are assayed for their homozygous performance of plant development, indicating which proteins are essential for plant growth.

I. Identification of Essential *Arabidopsis thaliana* Nucleotide Sequences and Encoded Proteins Using *Ac/Ds* Transposon or T-DNA-Mediated Mutagenesis

As shown in the examples below, the essentiality of the nucleotide sequences described herein for normal plant growth and development, have been demonstrated for the first time in *Arabidopsis* using *Ac/Ds* transposon or T-DNA-mediated mutagenesis. Having established the essentiality of the function of the encoded proteins in *Arabidopsis thaliana* and having identified the nucleotide sequences encoding these essential proteins, the inventors thereby provide an important and sought after tool for new herbicide development.

Arabidopsis insertional mutant lines segregating for seedling lethal mutations are identified as a first step in the identification of essential proteins. Starting with T2 seeds collected from single T1 plants containing T-DNA insertions in their genomes, those lines segregating homozygous seedling lethal seedlings are identified. Ds transposon insertion lines are produced as described in Sundaresan *et al.* (1995) (Genes and Dev., 9:1797-1810), incorporated herein by reference. Starting with F3 or F4 seeds collected from single F2 or F3 kanamycin-resistant plants containing *Ds* insertions in their genomes (see Figure 3 of Sundaresan *et al.* (1995) (Genes and Dev., 9:1797-1810), those lines segregating homozygous seedling lethal seedlings are identified. These lines are found by placing seeds onto minimal plant growth media, which contains the fungicides benomyl and maxim, and screening for inviable seedlings after 7 and 14 days in the light at room temperature. Inviable phenotypes include altered pigmentation or altered morphology. These phenotypes are observed either on plates directly or in soil following transplantation of seedlings.

Essential genes are also identified through the isolation of lethal mutants blocked in early development. Examples of lethal mutants include those blocked in the formation of the male or female gametes or embryo. Gametophytic mutants are found by examining T1 insertion lines for the presence of 50% aborted pollen grains or ovules. Embryo defective mutants produce 25% defective seeds following self-pollination of T1 plants (see Errampalli *et al.* 1991, Plant Cell 3:149-157; Castle *et al.* 1993, Mol Gen Genet 241:504-514).

When a line is identified as segregating a seedling lethal or an embryo defective phenotype, it is determined if the resistance marker in the *Ds* transposon or T-DNA insertion co-segregates with the lethality (Errampalli *et al.* (1991) The Plant Cell, 3:149-157). Cosegregation analysis is done by placing the seeds on media containing the selective agent and scoring the seedlings for resistance or sensitivity to the agent. Examples of selective agents used are kanamycin, hygromycin, or phosphinothricin. About 35 resistant seedlings are transplanted to soil and their progeny are examined for the segregation of the seedling lethal. In the case in which the *Ds* transposon or T-DNA insertion disrupts an essential gene, there is co-segregation of the resistance phenotype and the seedling lethal or embryo defective phenotype in every plant. Therefore, in such a case, all resistant plants segregate a seedling lethal or embryo defective phenotype in the next generation; this result indicates that each of the resistant plants is heterozygous for the mutation and hemizygous for the T-DNA insert causing the mutation.

For the *Arabidopsis* lines showing co-segregation of the transposon-encoded or T-DNA-encoded resistance marker and the lethal phenotype, PCR-based molecular approaches such as, TAIL-PCR (Liu *et al.* (1995) Plant J., 8:457-463; Liu and Whittier (1995), Genomics, 25:674-681), TAIL2k, vectorette PCR (Riley *et al.* (1990) Nucleic Acids Research, 18: 2887-2890)), or the GenomeWalker™ kit (CLONTECH Laboratories, Inc., Palo Alto, CA), may be used to directly amplify the plant DNA fragments flanking the transposon or T-DNA. Each of these techniques utilizes the known sequence of the transposon or T-DNA, and can be used to recover small (less than 5 kb) fragments directly adjacent to the insertion. PCR products are isolated and their DNA sequence is determined.

Alternatively, plasmid rescue may be used to isolate the plant DNA/T-DNA border fragments. Southern blot analysis may be performed as an initial step in the characterization of the molecular nature of each insertion. Southern blots are done with genomic DNA isolated from heterozygotes and using probes capable of hybridizing with the T-DNA vector DNA. Using the results of the Southern analysis, appropriate restriction enzymes are chosen to perform plasmid rescue in order to molecularly clone *Arabidopsis thaliana* genomic DNA flanking one or both sides of the T-DNA insertion. Plasmids obtained in this manner are analyzed by restriction enzyme digestion to sort the plasmids into classes based on their digestion pattern. For each class of plasmid clone, the DNA sequence is determined.

The resulting sequences, obtained by any of the above outlined approaches, are analyzed for the presence of non-*Ds* transposon and non-T-DNA vector sequences, as appropriate. When such sequences are found, they are used to search DNA and protein databases using the BLAST and BLAST2 programs (Altschul *et al.* (1990) *J Mol. Biol.* 215: 403-410; Altschul *et al.* (1997) *Nucleic Acid Res.* 25:3389-3402, both incorporated herein by reference). Additional genomic and cDNA sequences for each gene are identified by standard molecular biology procedures.

II. Recombinant Production Of Essential Proteins And Uses Thereof

For recombinant production of a protein of the invention in a host organism, a nucleotide sequence encoding the protein is inserted into an expression cassette designed for the chosen host and introduced into the host where it is recombinantly produced. The choice of the specific regulatory sequences such as promoter, signal sequence, 5' and 3' untranslated sequence, and enhancer appropriate for the chosen host is within the level of the skill of the routineer in the art. The resultant molecule, containing the individual elements linking in the proper reading frame, is inserted into a vector capable of being transformed into the host cell. Suitable expression vectors and methods for recombinant production of proteins are well known for host organisms such as *E. coli*, yeast, and insect cells (see, *e.g.*, Lucknow and Summers, *Bio/Technol.* 6:47 (1988)). Additional suitable expression vectors are baculovirus expression vectors, *e.g.*, those derived from the genome of *Autographica californica* nuclear polyhedrosis virus (AcMNPV). A preferred baculovirus/insect system is PVL1392(3) used to transfect *Spodoptera frugiperda* SF9 cells (ATCC) in the presence of linear *Autographica californica* baculovirus DNA (Phramingen, San Diego, CA). The resulting virus is used to infect HighFive *Tricoplusia ni* cells (Invitrogen, La Jolla, CA).

Recombinantly produced proteins are isolated and purified using a variety of standard techniques. The actual techniques used vary depending upon the host organism used, whether the protein is designed for secretion, and other such factors. Such techniques are well known to the skilled artisan (*see, e.g.* chapter 16 of Ausubel, F. *et al.*, "Current Protocols in Molecular Biology", pub. by John Wiley & Sons, Inc. (1994).

III. Assays For Characterizing The Essential Proteins

The recombinantly produced proteins described herein are useful for a variety of purposes. For example, they can be used in *in vitro* assays to screen known herbicidal chemicals whose target has not been identified to determine if they inhibit protein activity. Such *in vitro* assays may also be used as more general screens to identify chemicals that inhibit such protein activity and that are therefore novel herbicide candidates. Recombinantly produced proteins may also be used to elucidate the complex structure of these molecules and to further characterize their association with known inhibitors in order to rationally design new inhibitory herbicides. Alternatively, the recombinant protein can be used to isolate antibodies or peptides that modulate the activity and are useful in transgenic solutions.

IV. *In vitro* Inhibitor Assay: Discovery of Small Molecule Ligands That Interact with Essential Proteins Of Unknown Biochemical Function

Once a protein has been identified as a potential herbicide target based on its essentiality for normal plant growth and viability, a next step is to develop an assay that allows screening large number of chemicals to determine which ones interact with the protein. Although it is straightforward to develop assays for proteins of known function, developing assays with proteins of unknown functions can be more difficult.

To address this issue, novel technologies are used that can detect interactions between a protein and a compound without knowing the biological function of the protein. A short description of three methods is presented, including fluorescence correlation spectroscopy, surface-enhanced laser desorption/ionization, and biacore technologies.

Fluorescence Correlation Spectroscopy (FCS) theory was developed in 1972 but it is only in recent years that the technology to perform FCS became available (Madge *et al.* (1972) Phys. Rev. Lett., 29: 705-708; Maiti *et al.* (1997) Proc. Natl. Acad. Sci. USA, 94: 11753-11757). FCS measures the average diffusion rate of a fluorescent molecule within a small sample volume. The sample size can be as low as 10^3 fluorescent molecules and the sample volume as low as the cytoplasm of a single bacterium. The diffusion rate is a function of the mass of the molecule and decreases as the mass increases. FCS can therefore be applied to protein-ligand interaction analysis by measuring the change in mass and therefore in diffusion rate of a molecule upon binding. In a typical experiment, the target to be analyzed is expressed as a recombinant protein with a sequence tag, such as a poly-histidine

sequence, inserted at the N or C-terminus. The expression takes place in *E. coli*, yeast or insect cells. The protein is purified by chromatography. For example, the poly-histidine tag can be used to bind the expressed protein to a metal chelate column such as Ni2+ chelated on iminodiacetic acid agarose. The protein is then labeled with a fluorescent tag such as carboxytetramethylrhodamine or BODIPY® (Molecular Probes, Eugene, OR). The protein is then exposed in solution to the potential ligand, and its diffusion rate is determined by FCS using instrumentation available from Carl Zeiss, Inc. (Thornwood, NY). Ligand binding is determined by changes in the diffusion rate of the protein.

Surface-Enhanced Laser Desorption/Ionization (SELDI) was invented by Hutchens and Yip during the late 1980's (Hutchens and Yip (1993) Rapid Commun. Mass Spectrom. 7: 576-580). When coupled to a time-of-flight mass spectrometer (TOF), SELDI provides a mean to rapidly analyze molecules retained on a chip. It can be applied to ligand-protein interaction analysis by covalently binding the target protein on the chip and analyze by MS the small molecules that bind to this protein (Worrall *et al.* (1998) Anal. Biochem. 70: 750-756). In a typical experiment, the target to be analyzed is expressed as described for FCS. The purified protein is then used in the assay without further preparation. It is bound to the SELDI chip either by utilizing the poly-histidine tag or by other interaction such as ion exchange or hydrophobic interaction. The chip thus prepared is then exposed to the potential ligand via, for example, a delivery system capable to pipette the ligands in a sequential manner (autosampler). The chip is then submitted to washes of increasing stringency, for example a series of washes with buffer solutions containing an increasing ionic strength. After each wash, the bound material is analyzed by submitting the chip to SELDI-TOF. Ligands that specifically bind the target will be identified by the stringency of the wash needed to elute them.

Biacore relies on changes in the refractive index at the surface layer upon binding of a ligand to a protein immobilized on the layer. In this system, a collection of small ligands is injected sequentially in a 2-5 microlitre cell with the immobilized protein. Binding is detected by surface plasmon resonance (SPR) by recording laser light refracting from the surface. In general, the refractive index change for a given change of mass concentration at the surface layer, is practically the same for all proteins and peptides, allowing a single method to be applicable for any protein (Liedberg *et al.* (1983) Sensors Actuators 4: 299-304; Malmquist (1993) Nature, 361: 186-187). In a typical experiment, the target to be analyzed is expressed

as described for FCS. The purified protein is then used in the assay without further preparation. It is bound to the Biacore chip either by utilizing the poly-histidine tag or by other interaction such as ion exchange or hydrophobic interaction. The chip thus prepared is then exposed to the potential ligand via the delivery system incorporated in the instruments sold by Biacore (Uppsala, Sweden) to pipette the ligands in a sequential manner (autosampler). The SPR signal on the chip is recorded and changes in the refractive index indicate an interaction between the immobilized target and the ligand. Analysis of the signal kinetics on rate and off rate allows the discrimination between non-specific and specific interaction.

Another assay for small molecule ligands that interact with a polypeptide is an inhibitor assay. For example, such an inhibitor assay useful for identifying inhibitors of the products of essential plant nucleic acid sequences, such as the essential *Arabidopsis* proteins described herein, comprises the steps of:

a) reacting an essential *Arabidopsis* protein described herein and a substrate thereof in the presence of a suspected inhibitor of the protein's function;

b) comparing the rate of enzymatic activity of the protein in the presence of the suspected inhibitor to the rate of enzymatic activity under the same conditions in the absence of the suspected inhibitor; and

c) determining whether the suspected inhibitor inhibits the essential *Arabidopsis* protein.

For example, the inhibitory effect on the activity of a hereindescribed essential *Arabidopsis* protein, may be determined by a reduction or complete inhibition of protein activity in the assay. Such a determination may be made by comparing, in the presence and absence of the candidate inhibitor, the amount of substrate used or intermediate or product made during the reaction.

V. Production of peptides

Phage particles displaying diverse peptide libraries permits rapid library construction, affinity selection, amplification and selection of ligands directed against an essential protein (H.B. Lowman, *Annu. Rev. Biophys. Biomol. Struct.* 26, 401-424 (1997)). Structural analysis of these selectants can provide new information about ligand-target molecule interactions and

then in the process also provide a novel molecule that can enable the development of new herbicides based upon these peptides as leads.

VI. *In Vivo* Inhibitor Assay

In one embodiment, a suspected herbicide, for example identified by *in vitro* screening, is applied to plants at various concentrations. The suspected herbicide is preferably sprayed on the plants. After application of the suspected herbicide, its effect on the plants, for example death or suppression of growth is recorded.

In another embodiment, an *in vivo* screening assay for inhibitors of the activity of a hereindescribed essential protein uses transgenic plants, plant tissue, plant seeds or plant cells capable of overexpressing a nucleotide sequence disclosed herein that encodes an essential protein, wherein the essential protein is enzymatically active in the transgenic plants, plant tissue, plant seeds or plant cells. A chemical is then applied to the transgenic plants, plant tissue, plant seeds or plant cells and to the isogenic non-transgenic plants, plant tissue, plant seeds or plant cells, and the growth or viability of the transgenic and non-transformed plants, plant tissue, plant seeds or plant cells are determined after application of the chemical and compared. Compounds capable of inhibiting the growth of the non-transgenic plants, but not affecting the growth of the transgenic plants are selected as specific inhibitors of the essential protein's activity.

The invention will be further described by reference to the following detailed examples. These examples are provided for purposes of illustration only, and are not intended to be limiting unless otherwise specified.

EXAMPLES

Standard recombinant DNA and molecular cloning techniques used here are well known in the art and are described by J. Sambrook, *et al.*, Molecular Cloning: *A Laboratory Manual, 3d Ed.,* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press (2001); by T.J. Silhavy, M.L. Berman, and L.W. Enquist, *Experiments with Gene Fusions*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1984) and by Ausubel, F.M. *et al., Current Protocols in Molecular Biology*, New York, John Wiley and Sons Inc., (1988), Reiter, *et al., Methods in Arabidopsis Research*, World Scientific Press (1992), and Schultz *et al., Plant Molecular Biology Manual*, Kluwer Academic Publishers (1998). These references describe

the standard techniques used for all steps in tagging and cloning genes from *Ac/Ds* transposon or T-DNA mutagenized populations of *Arabidopsis*: plant infection and transformation; screening for the identification of seedling mutants; and cosegregation analysis. *Ds* transposon insertion lines produced as described in Sundaresan *et al.* (1995) Genes and Dev., 9:1797-1810) are used in these experiments. T-DNA lines are generated using vacuum infiltration or floral dip methods (Bechtold *et al.* (1993) C. R. Acad. Sci. Paris, 316:1194-1199; Clough and Bent (1998) Plant J., 16:735-743; Desfeux *et al.* (2000) Plant Physiol., 123:895-904).

Example 1: Identification of *Arabidopsis* Mutants with Lethal Phenotypes

Essential genes are identified through the isolation of lethal mutants blocked in early development. Examples of lethal mutants include those blocked in the formation of the male or female gametes, embryo, or resulting seedling. Gametophytic mutants are found by examining insertion lines for the presence of 50% aborted pollen grains or ovules. Embryo defective lethal mutants usually produce 25% defective seeds following self-pollination of plants heterozygous for an insertion (see Errampalli *et al.* 1991, Plant Cell 3:149-157; Castle *et al.* 1993, Mol Gen Genet 241:504-514). Seedling lethal mutants usually segregate 25% seedlings that exhibit a lethal phenotype.

Example 2: Cosegregation Analysis for Lines with Lethal Phenotypes

The linkage of the mutation to the *Ds* or T-DNA insertion is established after identifying a transformed line segregating for a lethal phenotype of interest. A line segregating with a single functional insert will segregate for resistance in the ratio of about 2:1 (resistant: sensitive) to the selectable marker. In the case of an embryo defective mutant, one-quarter of the progeny of a plant heterozygous for an insertion will fail to germinate due to embryo lethality, resulting in a reduction of the normal 3:1 ratio to 2:1. In the case of a seedling lethal mutant, the seedlings with a mutant phenotype are excluded in the calculation of this ratio. Each of the resistant progeny is therefore heterozygous for the mutation if the *Ds* or T-DNA insertion is causing the mutant phenotype. To establish cosegregation of the insertion and the mutant phenotype, about 30 resistant progeny are transplanted to soil and each plant is shown to segregate the 25% progeny with a lethal phenotype by the appropriate screening of embryo or seedlings. When all resistant plants segregate the lethal phenotype,

there is cosegregation of the insertion and the lethal mutation and the line is designated as "tagged."

Example 3: T-DNA Border Isolation by Plasmid Rescue

The plasmid rescue technique is used to molecularly clone *Arabidopsis* flanking DNA from one or both sides of the T-DNA insertion(s). *Arabidopsis* genomic DNA is isolated as described by Reiter *et al.* in *Methods in Arabidopsis Research*, World Scientific Press (1992). Genomic DNA is digested with a restriction endonuclease and ligated overnight. After ligation, the DNA is transformed into competent *E. coli* strain XL-1 Blue, DH10B, DH5 alpha, or the like, and colonies are selected on semi-solid medium containing ampicillin. Resistant colonies are picked into liquid medium with ampicillin and grown overnight. Plasmid DNA is isolated and digested with the rescue enzyme and analyzed on agarose gels containing ethidium bromide for visualization. Plasmids that represent different size classes are sequenced using primers that flank the plant DNA portion of the rescue element and the sequence is analyzed to determine what portion is plant DNA and what gene has been disrupted. The plasmid rescue is validated via PCR of template genomic DNA from a heterozygote for the insertion mutation. The experiment uses a primer anchored in the predicted flanking sequence and a primer in the T-DNA insertion. Finding a PCR product of the appropriate size, based on the sequence of the plasmid rescue clone confirms a valid rescue. Alternatively, Southern blot analysis with a probe that detects the relevant region of *Arabidopsis* DNA in genomic DNA from a heterozygote for the insertion mutation can be used to confirm the plasmid rescue results.

Example 4: Transposon or T-DNA Border Isolation by TAIL-PCR

Arabidopsis genomic DNA is isolated according to Reiter *et al.* in Methods in *Arabidopsis* Research, World Scientific Press (1992) or using the Nucleon PhytoPure™ Plant DNA isolation kit (Amersham International plc, Buckinghamshire, England) or the Puregene DNA isolation kit (Gentra Systems, Minneapolis, MN). Fragments of genomic DNA flanking the borders of the transposon or T-DNA are isolated using the TAIL-PCR technique (Liu *et al.* (1995) Plant J., 8:457-463; Liu and Whittier (1995), Genomics, 25:674-681). Three sets of 12 TAIL-PCR reactions, referred to as the primary, secondary and tertiary reactions, are performed. In each reaction, one arbitrary degenerate primer and one transposon-specific or

T-DNA-specific primer are used. The arbitrary degenerate primer is chosen from among seven primers, LWAD1, CA50, CA51, CA52, CA53, CA54, and CA55 (Table 1), which are used to prime the genomic DNA flanking the insertion. Alternatively, less than 12 TAIL-PCR reactions are done using fewer arbitrary degenerate primers. These degenerate primers are used in combination with two sets of three, nested, transposon-specific primers (Table 2) or T-DNA-specific primers (Table 3). The transposon-specific primers are homologous to regions of the *Ds* elements that lie at the outermost ends of the transposons, DS5 at the 5' end (primers 5A, 5B, and 5C) and DS3 at the 3' end (primers 3A, 3B, and 3C). The T-DNA-specific primers are homologous to regions of the T-DNA that lie in the borders of the T-DNAs. For the pCSA104 and pDAP101 T-DNAs, right borders are recovered with CA66 (primary primer), CA67 (secondary primer), and CA68 (tertiary primer) and left borders are recovered with JM33 (tertiary primer); JM34 (secondary primer); and JM35 (primary primer). For the pCSA110 T-DNA, right borders are recovered with QRB1 (primary primer), QRB2 (secondary primer), and QRB3 (tertiary primer) and left borders are recovered with JM33 (tertiary primer); JM34 (secondary primer); and JM35 (primary primer). For the pPCVICEn4HPT (Hayashi *et al.* (1992), Science, 258:1350-1353) and pSKI015 (Weigel *et al.* (2000) Plant Physiol. 122:1003-1014) T-DNAs, left borders are recovered with SKI1 (primary primer), SKI2 (secondary primer), and SKI3 (tertiary primer). When the degenerate and nested primer pairs are used in a series of low and high-stringency PCR amplifications, as described in the TAIL-PCR protocol (Liu and Whittier (1995), Genomics, 25:674-681), DNA fragments are produced that correspond to the genomic DNA that is directly adjacent to the transposon or T-DNA insertion. The nucleic acid sequences of the PCR products from the tertiary TAIL-PCR reactions are then determined by standard molecular biology techniques. The resulting sequences are analyzed for the presence of non-*Ds* transposon or non-T-DNA vector sequence.

To confirm the integrity of the resultant products, PCR primers specific to the flanking genomic region are designed and used in conjunction with the tertiary nested primer in a PCR reaction, to confirm the transposon or T-DNA insertion point within the genomic DNA. Finding a PCR product of the appropriate size, based on the sequence of the TAIL-PCR clone confirms a valid rescue.

Table 1: Arbitrary Degenerate Primers

SEQ ID NO:	Primer	Degen.	Primer Sequence
101	LWAD1	1026	ngt tgw gna twt sgw gnt
102	CA50	128	ngt cga swg ana wga a
103	CA51	128	tgw gna gsa nca sag a
104	CA52	128	agw gna gwa nca wag g
105	CA53	256	stt gnt ast nct ntg c
106	CA54	64	ntc gas twt sgw gtt
107	CA55	256	wgt gna gwa nca nag a

Table 2: Nested Primers For *Ds* Lines

SEQ ID NO:	Primer	Primer Sequence
108	5A	actagctctaccgtttccgtttccgtttac
109	5B	ttacctcgggttcgaaatcgatcgggataa
110	5C	aaaatcggttatacgataacggtcggtacggga
111	3A	gggtcttgcggatctgaatatatgttttcatgtgtg
112	3B	taccgaagaaaaataccggttcccgtccgatttcgac
113	3C	ggatcgtatcggttttcgattaccgtatttatcc

Table 3: Nested Primers For T-DNA Lines

SEQ ID NO:	Primer	Primer Sequence
114	CA66	att agg cac ccc agg ctt tac act tta tg
115	CA67	gta tgt tgt gtg gaa ttg tga gcg gat aac
116	CA68	taa caa ttt cac aca gga aac agc tat gac
117	JM33	tag cat ctg aat ttc ata acc aat ctc gat aca c
118	JM34	gct tcc tat tat atc ttc cca aat tac caa tac a
119	JM35	gcc ttt tca gaa atg gat aaa tag cct tgc ttc c
120	QRB1	caa act agg ata aat tat cgc gcg cgg tgt ca
121	QRB2	ggt gtc atc tat gtt act aga tcg gga att ga
122	QRB3	cgc cat ggc ata tgc tag cat gca taa ttc
123	SKI1	aat tgg taa tta ctc ttt ctt ttc ctc cat att ga
124	SKI2	ata ttg acc atc ata ctc att gct gat cca t
125	SKI3	tga tcc atg tag att tcc cgg aca tga a

Example 5: Transposon or T-DNA Border Isolation by TAIL2k PCR

Arabidopsis genomic DNA is isolated according to Reiter *et al.* in Methods in *Arabidopsis* Research, World Scientific Press (1992) or using the Nucleon PhytoPure™ Plant DNA isolation kit (Amersham International plc, Buckinghamshire, England) or the Puregene DNA isolation kit (Gentra Systems, Minneapolis, MN). Fragments of genomic DNA flanking the borders of the transposon or T-DNA are isolated using the TAIL2k PCR technique. Two sets of 12 TAIL-PCR reactions, referred to as the primary and secondary reactions, are performed. In each reaction, one arbitrary degenerate primer and one transposon-specific or T-DNA-specific primer are used. The arbitrary degenerate primer is selected from among six primers; CA50, CA51, CA52, CA53, CA54, and CA55 (Table 1), which are used to prime the genomic DNA flanking the insertion. Alternatively, less than 12 TAIL-PCR reactions are done using fewer arbitrary degenerate primers. These degenerate primers are used in combination with two sets of two, nested, transposon-specific primers (Table 2) or T-DNA-specific primers (Table 3). The transposon-specific primers are homologous to regions of the *Ds* elements that lie at the outermost ends of the transposons, DS5 at the 5' end (primers 5A, 5B, and 5C) and DS3 at the 3' end (primers 3A, 3B, and 3C). The T-DNA-specific primers are homologous to regions of the T-DNA that lie in the borders of the T-DNAs. For the pCSA104 and pDAP101 T-DNAs, right borders are recovered with CA66 (primary primer), CA67 (secondary primer), and CA68 (sequencing primer) and left borders are recovered with JM33 (sequencing primer), JM34 (secondary primer), and JM35 (primary primer). Primers CA66, CA67, and CA68 are also known as RB1, RB2, and RB3, respectively. Primers JM35, JM34, and JM33 are also known as LB1, LB2, and LB3, respectively. For the pCSA110 T-DNA, right borders are recovered with QRB1 (primary primer), QRB2 (secondary primer), and QRB3 (sequencing primer) and left borders are recovered with JM33 (sequencing primer); JM34 (secondary primer); and JM35 (primary primer). For the pPCVICEn4HPT (Hayashi *et al.* (1992), Science, 258:1350-1353) and pSKI015 (Weigel *et al.* (2000) Plant Physiol. 122:1003-1014) T-DNAs, left borders are recovered with SKI1 (primary primer), SKI2 (secondary primer), and SKI3 (sequencing primer). When the degenerate and nested primer pairs are used in a series of low and high-stringency PCR amplifications, as described in the TAIL-PCR protocol (Liu and Whittier (1995), Genomics, 25:674-681), DNA fragments are produced that correspond to the genomic DNA that is directly adjacent to the transposon

or T-DNA insertion. TAIL2k-PCR differs from the original TAIL-PCR protocol by the elimination of the tertiary PCR and modification of the secondary PCR. The cycling conditions used in the secondary reaction are modified to include 5 high annealing temperature cycles (64 degrees C) at the beginning, three additional so-called super cycles, and five additional low annealing temperature cycles (44 degrees C) at the end of the reaction. The melting and extension times are the same as all other TAIL-PCR reactions. Additionally, the reaction volume is increased to 40 microliters. The nucleic acid sequences of the PCR products from the secondary TAIL2k-PCR reactions are then determined by standard molecular biology techniques. The resulting sequences are analyzed for the presence of non-*Ds* transposon or non-T-DNA vector sequence.

To confirm the integrity of the resultant products, PCR primers specific to the flanking genomic region are designed and used in conjunction with the tertiary nested primer in a PCR reaction, to confirm the transposon or T-DNA insertion point within the genomic DNA. Finding a PCR product of the appropriate size, based on the sequence of the TAIL2k-PCR sequencing result confirms a valid rescue.

Example 6: Identification of Both Borders of a T-DNA or *Ds* Insertion

If the results of border rescue provide information on only one of the two borders for an insertion in a given line, additional experiments are performed to identify the second border. These experiments are necessary to show that a single gene has been disrupted in a given line. In some cases, an insertion can affect more than a single gene due to a chromosomal deletion or rearrangement. In those cases, additional experiments are required to identify which of the affected genes is responsible for the lethal phenotype.

When both borders of an insertion are not recovered, primers are designed to isolate a PCR product that will provide information on the location of the missing border. Three primers are chosen in *Arabidopsis* genomic DNA on the opposite side of the insertion about one, two, and five kb away from the insertion point; the primers point towards the expected second border. Long PCR conditions (Advantage 2, Clontech) are then employed following the manufacturer's directions to amplify the relevant region from genomic DNA isolated from a heterozygote for the lethal mutation. PCR reactions are performed using appropriate pairs of genomic and T-DNA or *Ds* border primers. Finding a PCR product of the appropriate size,

based on the sequence of the TAIL-PCR clone confirms a valid rescue of the second border. In some cases, the PCR product is directly sequenced to determine the exact insertion point.

If the second border is not recovered with this method, an additional set of PCR reactions are preformed. In these experiments, the genomic primers are paired with a series of internal T-DNA or *Ds* primers designed at about one kb intervals in both orientations across the entire T-DNA or *Ds* vector sequence. Finding a PCR product of the appropriate size, based on the sequence of the TAIL-PCR clone confirms a valid rescue of the second border. In some cases, the PCR product is directly sequenced to determine the exact insertion point. Any borders recovered with this approach are classified as abnormal because they lack the ends of the *Ds* transposon or the expected 24 bp T-DNA imperfect repeat characteristic of right and left borders.

Example 7: Identification of Insertion Points for Lines with Lethal Phenotypes

For each line with a lethal phenotype, the sequences of the borders of the insertion are determined and the insertion points in the *Arabidopsis* genome are deduced. For *Ds* insertion lines, PCR products are obtained from the Ds3 and Ds5 borders. For T-DNA lines, PCR products or plasmid rescue clones are obtained from left (LB), right (RB), or abnormal (AB) borders. These sequences are used in BLASTn searches against nucleotide databases (Altschul *et al.* (1990) J Mol. Biol. 215:403-410; Altschul *et al.* (1997) Nucleic Acids Res. 25:3389-3402). The results are summarized in Table 4. *Ds* line names begin with ET or GT; T-DNA line names are numbers. The insertion point (Insert Pt.) and the direction of the flanking sequence (Dir.) either up (U) or down (D) in the genome section is noted. Often, small deletions or duplications of genomic DNA accompany the insertion of a T-DNA or *Ds* transposon.

The gene that has been inactivated in a given line with a lethal phenotype is determined from the insertion points for that line. Often, the precise location of an ORF for a given gene is not known, but predictions are available in genome sections deposited in GenBank. The precise boundaries of that ORF is determined as described in Example 7.

Table 4: Insertion Points For Lines With Lethal Phenotypes

Gene	Line #	Border	Genome Section	Acc. #	Insert Pt.	Dir.
942	942	LB	K24G6	AB012242	33667	D

978	978	LB	F23N20	AC016972	58221	D
	978	LB	F23N20	AC016972	58301	U
3218	3218	LB	T8K14	AC007202	10500	D
	3218	LB	T8K14	AC007202	10540	U
4563	4563	LB	ATCHRII092	AC006438	25542	D
8794	8794	LB	F2J6	AC009526	45854	D
	8794	LB	F2J6	AC009526	45879	U
9106	9106	LB	T2J13	AL132967	78013	U
	9106	AB	T2J13	AL132967	77943	D
10708	10708	RB	F1I21	AC005687	40005	D
	10708	LB	F1I21	AC005687	40042	U
	70241	LB	F1I21	AC005687	40210	D
	70241	RB	F1I21	AC005687	40215	U
10844	10844	LB	F13F21	AC007504	60873	U
	10844	LB	F13F21	AC007504	60839	D
10951	10951	LB	MKP11	AB005238	20298	D
	10951	LB	MKP11	AB005238	20318	U
12935	12935	LB	ATCHRII150	AC005168	36510	D
	12935	LB	ATCHRII150	AC005168	36545	U
13823	11361	LB	T27G7	AC006932	78096	U
	11361	AB	T27G7	AC006932	78065	D
	13823	LB	T27G7	AC006932	78096	U
	13823	RB	T27G7	AC006932	77722	D
14519	14519	LB	ATCHRIV72	AL161576	50259	U
	14519	AB	ATCHRIV72	AL161576	50228	D
14610.1	14610.1	LB	F4P13	AC009325	55319	U
	14610.1	RB	F4P13	AC009325	55442	D
14891	14891	LB	ATCHRIV89	AL161593	11412	U
	14891	RB	ATCHRIV89	AL161593	11313	D
14986	14986	LB	K10D20	AP000410	51816	D
	14986	RB	K10D20	AP000410	54505	U
15377	15377	RB	F28G11	AC074025	19572	D
	15377	LB	F28G11	AC074025	19587	U
16219	16219	LB	MRO11	AB005244	51998	U
	16219	LB	MRO11	AB005244	51995	D
16547	16547	LB	ATCHRIV65	AL161565	80692	D
	16547	RB	ATCHRIV65	AL161565	80791	U
20933	20933	LB	ATCHRII146	AC004747	47678	D
	20933	LB	ATCHRII146	AC004747	47683	U
21455	21455	LB	ATCHRIV54	AL161554	105596	U
	21455	RB	ATCHRIV54	AL161554	105542	D
21878	21878	LB	T19F11	AC009918	19609	D
23915	23915	LB	ATCHRII008	AC005936	49629	D
	23915	LB	ATCHRII008	AC005936	49657	U
30945	30945	LB	ATCHRII192	AC004238	2411	D
	30945	LB	ATCHRII192	AC004238	2410	U

31895	31895	LB	MTI20	AB013396	52020	D
	31895	LB	MTI20	AB013396	52089	U
34269	34269	LB	T4O12	AC007396	92811	U
	34269	RB	T4O12	AC007396	92808	D
34540	34540	LB	T1G11	AC002376	41572	D
	34540	LB	T1G11	AC002376	41608	U
	72902	LB	T1G11	AC002376	41494	U
	72902	LB	T1G11	AC002376	41465	D
34555	34555	LB	T1F15	AC004393	42152	D
	54334	RB	T1F15	AC004393	41803	U
	54334	LB	T1F15	AC004393	41671	D
35154	35154	RB	MWD9	AB007651	45718	D
	35154	LB	MWD9	AB007651	45732	U
35438	35438	LB	MAL21	AP000383	25170	D
	35438	LB	MAL21	AP000383	25738	U
37351	37351	LB	F25C20	AC007296	52890	U
	37351	RB	F25C20	AC007296	52196	D
37389	37389	LB	F3F19	AC007357	45488	U
	37389	RB	F3F19	AC007357	45471	D
38108	38108	LB	ATCHRII150	AC005168	83430	D
	38108	RB	ATCHRII150	AC005168	83446	U
43301	43301	RB	T22D16	AL357612	57549	D
	43301	LB	T22D16	AL357612	57599	U
46250	46250	LB	F17A9	AC016827	74222	D
	46250	RB	F17A9	AC016827	74274	U
47050A	47050	LB	T23E18	AC009978	49445	D
	47050	RB	T23E18	AC009978	49475	U
52949A	52949	LB	K16H17	AB016884	34713	D
	52949	LB	K16H17	AB016884	34718	U
53210A	53210	RB	ATCHRII017	AC007167	92796	D
	53210	LB	ATCHRII017	AC007167	92942	U
	69121	LB	ATCHRII017	AC007167	94478	D
	69121	LB	ATCHRII017	AC007167	94502	U
55483	55483	RB	ATCHRII164	AC005727	71269	U
	55483	LB	ATCHRII164	AC005727	71258	D
58351A	58351	RB	MYH9	AB016893	42547	D
	58351	LB	MYH9	AB016893	42772	U
60944	60944	LB	F1B16	AC023754	89492	U
	60944	LB	F1B16	AC023754	89428	D
62837	62837	LB	T21J18	AL132963	70906	U
	62837	LB	T21J18	AL132963	70873	D
65310	65310	LB	T20H2	AC022472	8158	U
	65310	RB	T20H2	AC022472	8096	D
68181	68181	RB	F12A12	AL133314	38270	U
	68181	LB	F12A12	AL133314	38275	D
70913	70913	LB	T24H18	AL353013	5347	D

	70913	LB	T24H18	AL353013	5358	U
71067	71067	LB	F2E2	AC069252	63031	U
	71067	LB	F2E2	AC069252	62932	D
71654	71654	RB	MYA6	AB023046	71956	U
	71654	LB	MYA6	AB023046	71907	D
ET3172	ET3172	DS5	ATCHRIV4	AL161492	134442	U
ET3546	ET3546	DS3	ATCHRII115	AC006081	20874	D
	ET3546	DS5	ATCHRII115	AC006081	20973	U

Example 8: Identification of cDNAs for Essential Genes

A cDNA for a gene identified as essential is identified using a variety of approaches. This information enables the ORF for a given gene to be identified and used for other experiments including expression of the corresponding protein in heterologous systems.

If there is a full-length cDNA deposited in GenBank or published elsewhere, that sequence may be checked independently using methods described below. Alternatively, the sequence may be considered to be correct.

In some cases, there are published EST sequences that can be assembled to cover the entire ORF from start codon to stop codon. This sequence may be checked independently using methods described below or it may be considered to be correct.

Often part of the cDNA is published and this information can be used to identify the entire ORF. If the 5' end containing the start codon is known, 3' RACE is performed to identify the remainder of the cDNA. If the 3' end containing the stop codon is known, 5' RACE is performed to identify the remainder of the cDNA. If both the 5' and the 3' ends are known, but the sequence between the two ends of the cDNA is not known, PCR is performed with primers hybridizing to each end of the cDNA. In all three of these cases, PCR is performed using template DNA from a GeneRacer (Invitrogen) or a Marathon (Clontech) cDNA library prepared from RNA isolated from seedling tissue. A resulting PCR product is TA-cloned (Original TA-Cloning kit, Invitrogen) and sequenced.

If no part of the cDNA is published, the cDNA is identified by starting from gene model predictions in the annotation for genomic clones or elsewhere. To identify the ORF, primers are designed to the 5' and 3' ends of the predicted ORF. PCR is performed using template DNA from a cDNA library prepared from seedling tissue or the pFL61 *Arabidopsis* cDNA library (Minet *et al.* (1992) Plant J. 2: 417-422). The resulting PCR product is TA-cloned (Original TA-Cloning kit, Invitrogen) and sequenced. Alternatively, 5' and 3' RACE are performed with primers predicted by gene models to be in exons. PCR is performed using

template DNA from a GeneRacer (Invitrogen) or a Marathon (Clontech) cDNA library prepared from RNA isolated from seedling tissue. A resulting PCR product is TA-cloned (Original TA-Cloning kit, Invitrogen) and sequenced.

If the cDNA sequence is the same as the sequence predicted in the GenBank annotation, the experiments confirm for the first time the actual ORF. If the cDNA sequence is not the same as the sequence predicted in the GenBank annotation, the experiments identify for the first time the actual ORF. In some cases, more than one cDNA sequence is found for a given gene and both sequences are included in this application.

Example 9: Description of Essential Genes

The putative function of the protein encoded by each essential gene is determined from analysis of the ORF in each cDNA. Information from the relevant *Arabidopsis* genomic section deposited in GenBank is used as a starting point to explore the function of a given gene. This analysis also includes BLAST searches (Altschul *et al.* (1990) J. Mol. Biol. 215:403-410; Altschul *et al.* (1997) Nucleic Acids Res. 25:3389-3402) of sequence databases to identify similar proteins. Table 5 describes the putative functions for the essential genes discovered in this application.

Table 5: Putative Functions For Essential Genes

Gene	SEQ ID Nos:	Putative Function & Similar Genes	References
00942	1-2	similarity to disease resistance protein large gene family in *Arabidopsis* including disease resistance proteins RPP1-WsA,B&C; similar to tobacco TMV resistance protein N	Whitham, S. *et al.* (1994) Cell 78:1101-1115; Botella, M.A., *et al.* (1998) Plant Cell 10: 1847-1860
00978	3-4	unknown protein similar to *Arabidopsis* protein of unknown function (CAB87660) & ESTs from many plants	none
03218	5-6	AAA ATPase similar to *E. coli* FtsH cell division protein (P28691) that acts as an ATP-dependent metallopeptidase; homologs in many species	Schumann, W. (1999) FEMS Microbiol Rev 23:1-11; Langer, T. (2000) Trends Biochem Sci 2000 25:247-251

04563	7-8	unknown protein large gene family in *Arabidopsis* of unknown function proteins	none
08794	9-10	putative histidine decarboxylase similar to Brassica, tomato (tom92), and rice putative histidine decarboxylases	Picton, S *et al.* (1993) *Plant Mol Biol* 23:627-631; Watanabe, T *et al.* (1990) Trends Pharmacol Sci 11:363-367; Vaaler, G.L. & Snell, E.E. (1989) Biochemistry 28:7306-7313
09106	11-12	cytosolic 40S ribosomal protein S11-alpha	Browning, K.S. (1996) *Plant Mol Biol* 32:107-144; Gantt, J. S. & Thompson, M.D. (1990) *J. Biol Chem* 265:2763-2767
10708	13-14	cytoplasmic 60S ribosomal protein L3	Peltz, S.W. *et al.* (1999) *Mol Cell Biol* 19:384-391; Kim, Y. *et al.* (1990) *Gene* 93:177-182; Wickner, R.B *et al.* (1982) *Proc Natl Acad Sci USA* 79:4706-4708
10844	15-16	40S ribosomal protein S17-like	Gantt, J.S. & Thompson, M.D. (1990) *J Biol Chem* 265:2763-2767; Wiener, L. *et al.* (1988) *Nucleic Acids Res* 16:1233-1250
10951	17-18	phytoene synthase	Welsch, R. *et al.* (2000) *Planta* 211:846-854; Shewmaker, C.K. *et al.* (1999) *Plant J.* 20:401-412; Von Lintig, J. *et al.* (1997) *Plant J.* 12:625-634
12935	19-20	putative choline kinase similar to soybean choline kinase (T08815) and mouse & human choline/ethanolamine kinases	Monks, D.E. *et al.* (1996) Plant Physiol. 110:1197-1205; Bligny, R. *et al.* (1989) J Biol Chem. 264:4888-4895; Wharfe, J. & Harwood, J.L. (1979) Biochim Biophys Acta. 575:102-111
13823	21-22	magnesium protoporphyrin IX chelatase subunit D	Papenbrock, J. *et al.* (1997) Plant J. 12:981-990; Papenbrock, J. *et al.* (2000) Plant Physiol. 122:1161-1169; Luo, M. *et al.* (1999) Plant Mol Biol. 41:721-731; Jensen, P.E. *et al.* (1996) Mol. Gen. Genet. 250:383-394

14519	23-24	putative protein small gene family in *Arabidopsis* of unknown function proteins	none
14610.1	25-26	putative cell division control protein; similar to cdc48, AAA ATPase proteins similar to *S. pombe* AAA ATPase (CAB16902); *Arabidopsis* cdc48 homolog (P54609); cdc48/valosin-containing protein homologs from soybean, *Capsicum annuum*, rice, Dictylostelium; Drosophila smallminded	Frohlich, K.U. *et al.* (1991) J Cell Biol. 114:443-453; Feiler, H.S. *et al.* (1995) EMBO J. 14:5626-5637; Langer, T. (2000) Trends Biochem Sci 2000 25:247-251
14891	27-28	putative protein contains PFAM 02536 mTERF (mitochondrial transcription termination factor) domain; large gene family in *Arabidopsis* of unknown function proteins	Fernandez-Silva, P. *et al.* (1997) EMBO J 16:1066-1079
14986	29-30	ubiquitin isopeptidase T (aka ubiquitin-specific protease 14)	Wilkinson, K.D. *et al.* (1995) Biochemistry 34:14535-14546; Falquet, L. *et al.* (1995) FEBS Lett 376:233-237; Lindsey, D.F. *et al.* (1998) J Biol Chem 273:29178-29187
15377	31-32	putative formyl transferase similar to *B. napus* methionyl tRNA transformylase Fmt protein (AJ245479) & *B. rapa* S-locus protein 8 (AB022076)	Cui Y *et al.* (1999) Plant Cell. 11:2217-2231; Suzuki, G. *et al.* (1999) Genetics 153:391-400; Cusack S. (1999) Curr Opin Struct Biol. 9:66-73
16219	33-34	polyadenylation cleavage/specificity factor 100 kDa subunit (AF283277)	Bilger, A. *et al.* (1994) Genes Dev. 8:1106-1116; Bienroth, S. *et al.* (1993) EMBO J. 12:585-594; Jenny, A. *et al.* (1994) Mol Cell Biol. 14:8183-8190
16547	35-36	similarity to UV-induced protein Uvi31, *S. pombe*, G1381578 unknown function, but similar to Pectobacterium chrysanthemi SufE protein (AJ301654) involved in iron metabolism, *S. pombe* uvi31 protein of the BolA / YRBA family (Q12238), Synechocystis hypothetical 17.7 KDA protein SLR1419 (P74523)	Kim, S.H. *et al.* (1997) Environ Mol Mutagen 30:72-81; Santos, J.M. *et al.* (1999) Mol Microbiol 32:789-798

20933	37-38	hypothetical protein contains WD40 repeats, similar to human CIAO 1 gene (O76071) & *S. cerevisiae* YDR267c (S70127)	Neer, E.J. *et al.* (1994) Nature 371:297-300; Johnstone, R.W. *et al.* (1998) J Biol Chem 273:10880-10887
21455	39-40	putative protein small gene family in *Arabidopsis* of unknown function proteins	none
21878	41-42	*Arabidopsis* digalactosyldiacylglycerol synthase (DGD1, AAD42378)	Dormann, P. *et al.* (1999) *Science* 284:2181-2184; Hartel, H. *et al.* (1997) *Plant Physiol* 115:1175-1184
23915	43-44	hypothetical protein contains PPR motifs, member of large gene family in *Arabidopsis*	Small, I.D. & Peeters, N. (2000) *Trends Biochem Sci* 25:46-47; Manthey, G.M. & McEwen, J.E. (1995) *EMBO J* 14:4031-4043; Barkan, A. *et al.* (1994) *EMBO J* 13:3170-3181
30945	45-46	unknown protein similar to rice hypothetical protein (BAB56056)	none
31895	47-48	similar to unknown protein small gene family in *Arabidopsis* of unknown function proteins	none
34269	49-50	unknown protein	none
34540	51-52	probable lipoate protein ligase B, similar to Mycobacterium LIPB gene (O32961) also similar to *S. pombe* putative pre-tRNA/pre-rRNA processing protein (T41635)	Reed, K.E. & Cronan, J.E. Jr. (1993) J Bacteriol 175:1325-1336; Chen, X.J. (1997) Mol. Gen. Genet. 255:341-349
34555	53-54	similar to Synechocystis hypothetical 41.9KD protein (P52640) similar to several prokaryotic proteins of unknown function including E. coli YJEQ (P39286)	none
35154	55-56	similar to unknown protein similar to human hypothetical protein (BAA91556), *S. cerevisiae* probable membrane protein YOR262w (S67159), & *S. pombe* ATP(GTP)-binding protein Fet5 (AAC49837)	Shpakovskii, G.V. & Lebedenko, E.N. (1997) Bioorg. Khim. 23:234-237
35438	57-58	unknown protein weak similarity to *Pennisetum ciliare* unknown function protein (AAK15504)	none

37351	59-60	strong similarity to obtusifoliol 14-alpha demethylase (CYP51; P93846) from Sorghum bicolor (also wheat & rice), member of the PF|00067 cytochrome P450 family	Kushiro, M. *et al.* (2001) Biochem Biophys Res Commun. 285:98-104; Bak *et al.* (1997) Plant J. 11:191-201; Grausem, B. (1995) Plant J. 7:761-770
37389	61-62	similar to human GLE1-like required for poly(A)+ RNA export (AAC25561)	Watkins, J.L. *et al.* (1998) Proc. Natl. Acad. Sci. U.S.A. 95:6779-6784; Murphy, R. & Wente, S.R. (1996) Nature 383:357-360
38108	63-64	*Arabidopsis* 4-(cytidine 5'-phospho)-2-C-methyl-D-erythritol kinase (aka ispE & 4-diphosphocytidyl-2-C-methyl-Derythritol kinase) (AF288615) similar to *E. coli* ychB (aka ispE) gene (P24209)	Rohdich, F. *et al.* (2000) Proc Natl Acad Sci U.S.A. 97:8251-8256; Luettgen, H. *et al.* (2000) Proc. Natl. Acad. Sci. U.S.A. 97:1062-1067; Lange, B.M. & Croteau, R. (1999) Proc Natl Acad Sci U.S.A. 96:13714-13719
43301	65-66	similar to hypothetical bacterial proteins, including *Pseudomonas aeruginosa* protein PA0292 (F83608) & *Lactococcus lactis* (AAK05795)	none
46250	67-68	hypothetical protein weak similarity to hypothetical proteins from *Arabidopsis* (AAG51506) and mouse (BAB23375)	none
47050A	69-70	unknown protein weak similarity to Botrytis cDNA (AL115827)	none
52949A	71-72	6-phosphogluconolactonase-like protein similar to 6-phosphogluconolactonases such as human (O95336), Brassica carinata (AAK50346), & *Mycobacterium tuberculosis* (devB, CAB09261)	Collard, F. *et al.* (1999) FEBS Lett. 459:223-226; Bauer, H.P. *et al.* (1983) Eur J Biochem. 133:163-168
53210A	73-74	putative heat shock protein in hsp90 family similar to rye hsp82 (S65776), Ipomoea nil hsp83 (P51819), chicken hsp90 beta (Q04619) and others	Felsheim, R.F. & Das, A. (1992) Plant Physiol. 100:1764-1771; Coates, A.R. *et al.* (1999) *Biotechnol Genet Eng Rev* 16:393-405; Milioni, D. & Hatzopoulos, P. (1997) *Plant Mol Biol* 35:955-961

55483	75-76	putative para-aminobenzoate synthase and glutamine amidotransferase, a bifunctional enzyme similar to *Streptomyces pristinaespiralis* papA (AAC44866), *E. coli* pabB (P05041) & pabA (P00903), and *Bacillus stearothermophilus* anthranilate synthase component I trpE (AAD33791)	Goncharoff, P. & Nichols, B.P. (1984) J Bacteriol. 159:57-62.; Roux, B. & Walsh, C.T. (1992) Biochemistry. 31:6904-6910; Kaplan, J.B. & Nichols, B.P. (1983) J Mol Biol 168:451-468
58351A	77-78	26S proteasome p55 protein-like similar to human 26S proteasome regulatory complex chain p55 (BAA19749), *S. cerevisiae* 26S proteasome regulatory complex chain RPN5 (S67695), and others	Saito, A. *et al.* (1997) *Gene* 203:241-250; Glickman, M.H. *et al.* (1998) Mol Cell Biol 18:3149-3162
60944	79-80	similar to Guillardia theta chloroplast 50S ribosomal protein L31 (O46917)	Yamaguchi, K. & Subramanian, A.R. (2000) *J Biol Chem*, 275:28466-28482
62837	81-82	AtClpC: regulatory subunit of Clp protease with ATPase activity (BAA82062)	Adam, Z. (2000) Biochimie 82:647-654; Sokolenko, A. *et al.* (1998) Planta 207:286-295; Nakabayashi, K. *et al.* (1999) Plant Cell Physiol. 40:504-514; Maurizi, M.R. *et al.* (1990) J Biol Chem. 265:12536-12545
65310	83-84	26S proteasome regulatory subunit S3, contains a PCI PF01399 domain similar to 26S proteasome regulatory subunit S3 from Nicotiana tabacum (P93768), carrot (Q06364), human (O43242), *S. cerevisiae* RPN3 (P40016), and others	Voges, D. *et al.* (1999) Ann Rev Biochem 68:1015-1068; Fu, H. *et al.* (1999) Mol Biol Rep 26:137-146; Fu, H. *et al.* (1999) Plant J 18:529-539; Kominami, K. *et al.* (1997) Mol Biol Cell 8:171-187
68181	85-86	small zinc finger-like protein TIM9 similar to mitochondrial import inner membrane translocase subunit TIM9 from several plants and *S. cerevisiae* (O74700)	Koehler, C.M. *et al.* (1998) EMBO J. 17:6477-6486; Tokatlidis, K. *et al.* (2000) Biochem Soc Trans 28:495-499
70913	87-88	*Arabidopsis* CCAAT binding protein/transcription factor Hap2a (CAA74048)	Edwards, D. *et al.* (1998) Plant Physiol 117:1015-1022; Albani, D. & Robert, L.S. (1995) *Gene* 167:209-213

71067	89-90	hypothetical protein gene family in *Arabidopsis* of unknown function proteins	none
71654	91-92	poly(A) binding protein-like	Hilson, P. *et al.* (1993) *Plant Physiol* 103:525-533; Belostotsky, D.A. & Meagher, R.B. (1993) *Proc Natl Acad Sci U.S.A.* 90:6686-6690; Gallie, D.R. (1998) *Gene* 216:1-11
ET3172	93-94	hypothetical protein small gene family in *Arabidopsis* (T47999 & T02193) of unknown function	none
ET3546	95-96	cdc27/nuc2-like protein, may contain TPR-repeat similar to human cdc27 (P30260), *S. pombe* nuc2 (P10505), *S. cerevisiae* cdc23 (P16522), and others	Hirano, T. *et al.* (1988) J. Cell Biol. 106:1171-1183; Chen, P.L. *et al* (1995) Cell Growth Differ. 6:199-210

Example 10: Expression of Recombinant Essential Proteins in *E. coli*

The coding region of each of the essential proteins, corresponding to cDNA clones of odd-numbered SEQ ID NO:1-96, is subcloned into an appropriate expression vector, and transformed into *E. coli* using the manufacturer's conditions. Specific examples include plasmids such as pBluescript (Stratagene, La Jolla, CA), pFLAG (International Biotechnologies, Inc., New Haven, CT), and pTrcHis (Invitrogen, La Jolla, CA). *E. coli* is cultured, and expression of the essential protein is confirmed. Recombinant protein is isolated using standard techniques.

Example 11: *In Vitro* Binding Assays

Recombinant protein for each of the essential genes described in this application is obtained, for example, according to Example 10. The protein is immobilized on chips appropriate for ligand binding assays using techniques that are well known in the art. The protein immobilized on the chip is exposed to sample compound in solution according to methods well know in the art. While the sample compound is in contact with the immobilized protein, measurements capable of detecting protein-ligand interactions are conducted. Examples of such measurements are SELDI, biacore and FCS, described above. Compounds

found to bind the protein are readily discovered in this fashion and are subjected to further characterization.

The above-disclosed embodiments are illustrative. This disclosure of the invention will place one skilled in the art in possession of many variations of the invention. All such obvious and foreseeable variations are intended to be encompassed by the present invention.

CLAIMS:

1. A method of identifying a herbicidal compound, comprising:
   a) combining a polypeptide comprising an amino acid sequence at least 90% identical to an amino acid sequence selected from the group consisting of the even numbered SEQ ID NOs:2-96 with a compound to be tested for the ability to bind to said polypeptide, under conditions conducive to binding;
   b) selecting a compound identified in (a) that binds to said polypeptide;
   c) applying a compound selected in (b) to a plant to test for herbicidal activity; and
   d) selecting a compound identified in (c) that has herbicidal activity.

2. The method according to claim 1, wherein said polypeptide comprises an amino acid sequence at least 95% identical to an amino acid sequence selected from the group consisting of the even numbered SEQ ID NOs:2-96.

3. The method according to claim 2, wherein said polypeptide comprises an amino acid sequence at least 99% identical to an amino acid sequence selected from the group consisting of the even numbered SEQ ID NOs:2-96.

4. The method according to claim 3, wherein said polypeptide comprises an amino acid sequence selected from the group consisting of the even numbered SEQ ID NOs:2-96.

5. A method of identifying a herbicidal compound, comprising:
   c) combining a polypeptide comprising an amino acid sequence at least 90% identical to an amino acid sequence selected from the group consisting of the even numbered SEQ ID NOs:2-96 with a compound to be tested for the ability to inhibit the activity of said polypeptide, under conditions conducive to inhibition;
   d) selecting a compound identified in (a) that inhibits the activity of said polypeptide;
   c) applying a compound selected in (b) to a plant to test for herbicidal activity; and
   d) selecting a compound identified in (c) that has herbicidal activity.

6. The method according to claim 5, wherein said polypeptide comprises an amino acid sequence at least 95% identical to an amino acid sequence selected from the group consisting of the even numbered SEQ ID NOs:2-96.

7. The method according to claim 6, wherein said polypeptide comprises an amino acid sequence at least 99% identical to an amino acid sequence selected from the group consisting of the even numbered SEQ ID NOs:2-96.

8. The method according to claim 7, wherein said polypeptide comprises an amino acid sequence selected from the group consisting of the even numbered SEQ ID NOs:2-96.

9. A method for killing or inhibiting the growth or viability of a plant, comprising applying to the plant a herbicidal compound identified according to the method of claim 1.

10. A method for killing or inhibiting the growth or viability of a plant, comprising applying to the plant a herbicidal compound identified according to the method of claim 5.

Lys Leu Glu Met Lys Asp Lys Tyr Glu Lys Leu Met Lys Lys Tyr Pro
210 215 220

Pro Pro Gln Trp Glu Phe Arg Tyr Ile Lys Gly Arg Arg Val Lys Val
225 230 235 240

Lys Ala Lys Gln Leu Asn Glu Leu Ser Glu Gly Glu Gly Gly Leu Ser
245 250 255

Ser Asp Glu Asp Lys Ile Asp Asn Glu Ile Glu Ser Glu Glu Glu Asp
260 265 270

Gly Glu Asp Leu Ser Glu Glu Glu Glu Asp Glu Lys Glu Leu Leu Gly
275 280 285

Gly Ser Gln Gly Gln Ile Thr Ser Arg Glu Pro Ser Leu Asp His Leu
290 295 300

Asp Ser Ser
305

<210> 41

<211> 2427

<212> DNA

<213> Arabidopsis thaliana

<220>

<221> CDS

<222> (1)..(2427)

<223> 21878

<400> 41

atg gta aag gaa act cta att cct ccg tca tct acg tca atg acg acc 48
Met Val Lys Glu Thr Leu Ile Pro Pro Ser Ser Thr Ser Met Thr Thr
1 5 10 15

gga aca tct tct tct tcg tct ctt tca atg acg tta tcc tca aca aac 96
Gly Thr Ser Ser Ser Ser Ser Leu Ser Met Thr Leu Ser Ser Thr Asn
20 25 30

gcg tta tcg ttt ttg tcg aaa gga tgg aga gag gta tgg gat tca gca 144
Ala Leu Ser Phe Leu Ser Lys Gly Trp Arg Glu Val Trp Asp Ser Ala

```
        35                  40                  45

gat gcg gat ttg cag ctg atg cga gac aga gct aac tct gtt aag aat      192
Asp Ala Asp Leu Gln Leu Met Arg Asp Arg Ala Asn Ser Val Lys Asn
    50                  55                  60

cta gca tca acg ttc gat aga gag atc gag aat ttc ctc aat aac tcg      240
Leu Ala Ser Thr Phe Asp Arg Glu Ile Glu Asn Phe Leu Asn Asn Ser
65                  70                  75                  80

gcg agg tct gcg ttt ccc gtt ggt tca cca tcg gcg tcg tct ttc tca      288
Ala Arg Ser Ala Phe Pro Val Gly Ser Pro Ser Ala Ser Ser Phe Ser
                85                  90                  95

aat gaa att ggt atc atg aag aag ctt cag ccg aag att tcg gag ttt      336
Asn Glu Ile Gly Ile Met Lys Lys Leu Gln Pro Lys Ile Ser Glu Phe
            100                 105                 110

cgt agg gtt tat tcg gcg ccg gag att agt cgc aag gtt atg gag aga      384
Arg Arg Val Tyr Ser Ala Pro Glu Ile Ser Arg Lys Val Met Glu Arg
        115                 120                 125

tgg gga cct gcg aga gcg aag ctt gga atg gat cta tcg gcg att aag      432
Trp Gly Pro Ala Arg Ala Lys Leu Gly Met Asp Leu Ser Ala Ile Lys
    130                 135                 140

aag gcg att gtg tct gag atg gaa ttg gat gag cgt cag gga gtt ttg      480
Lys Ala Ile Val Ser Glu Met Glu Leu Asp Glu Arg Gln Gly Val Leu
145                 150                 155                 160

gag atg agt aga ttg agg aga cgg cgt aat agt gat agg gtt agg ttt      528
Glu Met Ser Arg Leu Arg Arg Arg Arg Asn Ser Asp Arg Val Arg Phe
                165                 170                 175

acg gag ttt ttc gcg gag gct gag aga gat gga gaa gct tat ttc ggt      576
Thr Glu Phe Phe Ala Glu Ala Glu Arg Asp Gly Glu Ala Tyr Phe Gly
            180                 185                 190

gat tgg gaa ccg att agg tct ttg aag agt aga ttt aaa gag ttt gag      624
Asp Trp Glu Pro Ile Arg Ser Leu Lys Ser Arg Phe Lys Glu Phe Glu
        195                 200                 205

aaa cga agc tcg tta gaa ata ttg agt gga ttc aag aac agt gaa ttt      672
Lys Arg Ser Ser Leu Glu Ile Leu Ser Gly Phe Lys Asn Ser Glu Phe
    210                 215                 220

gtt gag aag ctc aaa acc agc ttt aaa tca att tac aaa gaa act gat      720
Val Glu Lys Leu Lys Thr Ser Phe Lys Ser Ile Tyr Lys Glu Thr Asp
225                 230                 235                 240

gag gct aag gat gtc cct ccg ttg gat gta cct gaa ctg ttg gca tgt      768
Glu Ala Lys Asp Val Pro Pro Leu Asp Val Pro Glu Leu Leu Ala Cys
                245                 250                 255

ttg gtt aga caa tct gaa cct ttt ctt gat cag att ggt gtt aga aag      816
Leu Val Arg Gln Ser Glu Pro Phe Leu Asp Gln Ile Gly Val Arg Lys
            260                 265                 270

gat aca tgt gac cga ata gta gaa agc ctt tgc aaa tgc aag agc caa      864
Asp Thr Cys Asp Arg Ile Val Glu Ser Leu Cys Lys Cys Lys Ser Gln
```

```
        275                 280                 285

caa ctt tgg cgt ctg cca tct gca caa gca tcc gat tta att gaa aat      912
Gln Leu Trp Arg Leu Pro Ser Ala Gln Ala Ser Asp Leu Ile Glu Asn
    290                 295                 300

gat aac cat gga gtt gat ttg gat atg agg ata gcc agt gtt ctt caa      960
Asp Asn His Gly Val Asp Leu Asp Met Arg Ile Ala Ser Val Leu Gln
305                 310                 315                 320

agc aca gga cac cat tat gat ggt ggg ttt tgg act gat ttt gtg aag     1008
Ser Thr Gly His His Tyr Asp Gly Gly Phe Trp Thr Asp Phe Val Lys
                325                 330                 335

cct gag aca ccg gaa aac aaa agg cat gtg gca att gtt aca aca gct     1056
Pro Glu Thr Pro Glu Asn Lys Arg His Val Ala Ile Val Thr Thr Ala
            340                 345                 350

agt ctt cct tgg atg acc gga aca gct gta aat ccg cta ttc aga gcg     1104
Ser Leu Pro Trp Met Thr Gly Thr Ala Val Asn Pro Leu Phe Arg Ala
        355                 360                 365

gcg tat ttg gca aaa gct gca aaa cag agt gtt act ctc gtg gtt cct     1152
Ala Tyr Leu Ala Lys Ala Ala Lys Gln Ser Val Thr Leu Val Val Pro
    370                 375                 380

tgg ctc tgc gaa tct gat caa gaa cta gtg tat cca aac aat ctc acc     1200
Trp Leu Cys Glu Ser Asp Gln Glu Leu Val Tyr Pro Asn Asn Leu Thr
385                 390                 395                 400

ttc agc tca cct gaa gaa caa gag agt tat ata cgt aaa tgg ttg gag     1248
Phe Ser Ser Pro Glu Glu Gln Glu Ser Tyr Ile Arg Lys Trp Leu Glu
                405                 410                 415

gaa agg att ggt ttc aag gct gat ttt aaa atc tcc ttt tac cca gga     1296
Glu Arg Ile Gly Phe Lys Ala Asp Phe Lys Ile Ser Phe Tyr Pro Gly
            420                 425                 430

aag ttt tca aaa gaa agg cgc agc ata ttt cct gct ggt gac act tct     1344
Lys Phe Ser Lys Glu Arg Arg Ser Ile Phe Pro Ala Gly Asp Thr Ser
        435                 440                 445

caa ttt ata tcg tca aaa gat gct gac att gct ata ctt gaa gaa cct     1392
Gln Phe Ile Ser Ser Lys Asp Ala Asp Ile Ala Ile Leu Glu Glu Pro
    450                 455                 460

gaa cat ctc aac tgg tat tat cac ggc aag cgt tgg act gat aaa ttc     1440
Glu His Leu Asn Trp Tyr Tyr His Gly Lys Arg Trp Thr Asp Lys Phe
465                 470                 475                 480

aac cat gtt gtt gga att gtc cac aca aac tac tta gag tac atc aag     1488
Asn His Val Val Gly Ile Val His Thr Asn Tyr Leu Glu Tyr Ile Lys
                485                 490                 495

agg gag aag aat gga gct ctt caa gca ttt ttt gtg aac cat gta aac     1536
Arg Glu Lys Asn Gly Ala Leu Gln Ala Phe Phe Val Asn His Val Asn
            500                 505                 510

aat tgg gtc aca cga gcg tat tgt gac aag gtt ctt cgc ctc tct gcg     1584
Asn Trp Val Thr Arg Ala Tyr Cys Asp Lys Val Leu Arg Leu Ser Ala
```

```
        515                 520                 525

gca aca caa gat tta cca aag tct gtt gta tgc aat gtc cat ggt gtc     1632
Ala Thr Gln Asp Leu Pro Lys Ser Val Val Cys Asn Val His Gly Val
    530                 535                 540

aat ccc aag ttc ctt atg att ggg gag aaa att gct gaa gag aga tcc     1680
Asn Pro Lys Phe Leu Met Ile Gly Glu Lys Ile Ala Glu Glu Arg Ser
545                 550                 555                 560

cgt ggt gaa caa gct ttc tca aaa ggt gca tac ttc tta gga aaa atg     1728
Arg Gly Glu Gln Ala Phe Ser Lys Gly Ala Tyr Phe Leu Gly Lys Met
                565                 570                 575

gtg tgg gct aaa gga tac aga gaa cta ata gat ctg atg gct aaa cac     1776
Val Trp Ala Lys Gly Tyr Arg Glu Leu Ile Asp Leu Met Ala Lys His
            580                 585                 590

aaa agc gaa ctt ggg agc ttc aat cta gat gta tat ggg aac ggt gaa     1824
Lys Ser Glu Leu Gly Ser Phe Asn Leu Asp Val Tyr Gly Asn Gly Glu
        595                 600                 605

gat gca gtc gag gtc caa cgt gca gca aag aaa cat gac ttg aat ctc     1872
Asp Ala Val Glu Val Gln Arg Ala Ala Lys Lys His Asp Leu Asn Leu
    610                 615                 620

aat ttc ctc aaa gga agg gac cac gct gac gat gct ctt cac aag tac     1920
Asn Phe Leu Lys Gly Arg Asp His Ala Asp Asp Ala Leu His Lys Tyr
625                 630                 635                 640

aaa gtg ttc ata aac ccc agc atc agc gat gtt cta tgc aca gca acc     1968
Lys Val Phe Ile Asn Pro Ser Ile Ser Asp Val Leu Cys Thr Ala Thr
                645                 650                 655

gca gaa gca cta gcc atg ggg aag ttt gtg gtg tgt gca gat cac cct     2016
Ala Glu Ala Leu Ala Met Gly Lys Phe Val Val Cys Ala Asp His Pro
            660                 665                 670

tca aac gaa ttc ttt aga tca ttc ccg aac tgc tta act tac aaa aca     2064
Ser Asn Glu Phe Phe Arg Ser Phe Pro Asn Cys Leu Thr Tyr Lys Thr
        675                 680                 685

tcc gaa gac ttt gtg tcc aaa gtg caa gaa gca atg acg aaa gag cca     2112
Ser Glu Asp Phe Val Ser Lys Val Gln Glu Ala Met Thr Lys Glu Pro
    690                 695                 700

cta cct ctc act cct gaa caa atg tac aat ctc tct tgg gaa gca gca     2160
Leu Pro Leu Thr Pro Glu Gln Met Tyr Asn Leu Ser Trp Glu Ala Ala
705                 710                 715                 720

aca cag agg ttc atg gag tat tca gat ctc gat aag atc tta aac aat     2208
Thr Gln Arg Phe Met Glu Tyr Ser Asp Leu Asp Lys Ile Leu Asn Asn
                725                 730                 735

gga gag gga gga agg aag atg cga aaa tca aga tcg gtt ccg agc ttt     2256
Gly Glu Gly Gly Arg Lys Met Arg Lys Ser Arg Ser Val Pro Ser Phe
            740                 745                 750

aac gag gtg gtc gat gga gga ttg gca ttc tca cac tat gtt cta aca     2304
Asn Glu Val Val Asp Gly Gly Leu Ala Phe Ser His Tyr Val Leu Thr
```